Research on the Factors Improving the Health of County Level in the United States

DATA5207: Data Analysis in the Social Sciences

Jian Gu (480480938)

Introduction

According to a study by the Institute for Health Metrics and Evaluation (IHME), the life expectancy of many county-level cities in the United States has not increased compared with the 1980s. Compared with the health status of cities, rural residents are facing a greater impact. Residents' health is not only determined by their physical fitness and income level, but by the combined effects of education, living habits and medical insurance.

"We can't be a healthy, thriving nation if we continue to leave entire communities and populations behind," said Richard Besser, M.D., Robert Wood Johnson Foundation (RWJF) president and CEO in 2018. In the United States, every community should use the health rankings of its county-level cities so that they can work together to find solutions in order to give all infants, children, and adults the same equal right to be healthy.

This project aims to formulate a theoretical explanation model that promotes the healthy development of cities by analyzing and determining various factors of the health of county-level urban residents in the United States. This report will use the Ordered Probit Regression model to analyze the 2018 County Health Rankings data in the Robert Wood Johnson Foundation project. Stepwise regression is used to incorporate various factors into the regression model to verify various health influencing factors. Finally, the opinions of medical reform and the emphasis on health education are put forward.

Literature and theory

Rita, Daniel and Sanjay (2019) conducted a longitudinal survey of American youth on the association of county-level social factors with individual smoking and drinking. The main predictors are three factors in the county-level city where you live: education, unemployment and per capita income. The project team adjusts the individual level covariates through the ordinary least squares (OLS) model, and then draws conclusions through the fixed effects (FE) model. Under the county-level socio-economic characteristics, place factors are the root cause of health behavior (smoking and drinking). Socioeconomic disadvantages at the regional level affect health outcomes.

Among socio-demographic peer groups, compare the results of smoking, motor vehicle crash death and obesity in county-level cities, and establish effective comparison groups to evaluate county-level public health performance. This will help local health departments share technology and meet future public health challenges. The counties with young, urban, and mid-to-high socioeconomic status groups have fallen by 60% or more compared to the national ranking of all three related results (Wallace, Sharfstein& Kaminsky, 2019).

In the analysis of the county-level mortality rates in the United States from 1980 to 2014, using a novel method, the analysis found that the 21 main causes of death and mortality rates in various counties in the United States changed significantly. The use of small-area models for county-level analysis has the potential

to provide novel insights into the temporal trends of mortality rates for specific diseases in the United States and their differences between geographic regions (Jama, 2016).

Using public databases from 2008 to 2012 to study 3,112 county-level cities in the United States, multi-factor analysis shows that poverty, unemployment, partisan voting, and the percentage of Hispanic and Native American Indians in a county are important causes of low health insurance coverage index (Stone& Boursaw, 2015).

From 1980 to 2014, cardiovascular disease was the leading cause of death in the United States. Through analysis of cardiovascular mortality in 3110 residential counties, the counties with the highest mortality extend from southeastern Oklahoma along the Mississippi River Basin to eastern Kentucky. There are basically several cardiovascular diseases in the southern region, including atrial fibrillation, aortic aneurysm and endocarditis. The areas with the lowest cardiovascular mortality are San Francisco, California, and central Colorado (Roth, 2017).

Between 2010 and 2017, the number of potential excess deaths between most rural and urban counties in the United States has increased, the gap in accidental injuries has decreased, and the gap in stroke has been relatively stable. The increase in the number of potential excess deaths due to accidental injuries has led to a narrowing of the already high percentage of excess deaths in non-core counties and small city counties (Garcia, 2019).

Over the past few decades, the US healthcare system has undergone tremendous changes. New and improved medicines, equipment, procedures, and imaging instruments have changed the mode of care and provided new places of care (NCHS, 2016). But this does not mean that all Americans have equal access. For example, white women are more likely to undergo outpatient surgery than women of other races (Salasky, 2014), and poor and disabled people are more likely to use the emergency department than people with other insurance, partly because they have less access to outpatient treatment (Medicaid and CHIP Payment and Access Commission [MACPAC], 2016). Despite the overall improvement in health status, there are still huge social gaps in many health indicators, most notably life expectancy and infant mortality (Singh, 2017).

The obvious differences between various health outcomes indicate that social determinants have fundamental significance in disease prevention and health promotion, so it is necessary to systematically and continuously monitor health inequality based on social factors. A multisectoral approach is needed to resolve the persistent and expanding health inequalities among Americans.

Data and methodology

The data used in this study comes from the centers for disease and disease control from the US Census Bureau, the data comes from different sources mainly including variables of health outcomes and related factors such as Length of Life, Quality of Life, Health Behaviors, Physical Environment and so on. There are lots of the factors in the data, as the factors and response are in different data sets, the data sets are firstly merged using the county-level indicator which is the combination of Federal Information Processing Standard (FIPS), State and County. Also, missing values are removed such as “NR” in the health outcome, and strings are converted into numeric values. The response used in the study is the rank of health outcome for counties and the factors are all in quartile scales which are ranged from 1 to 5 because all of them are ordered data, so the linear model is appropriate.

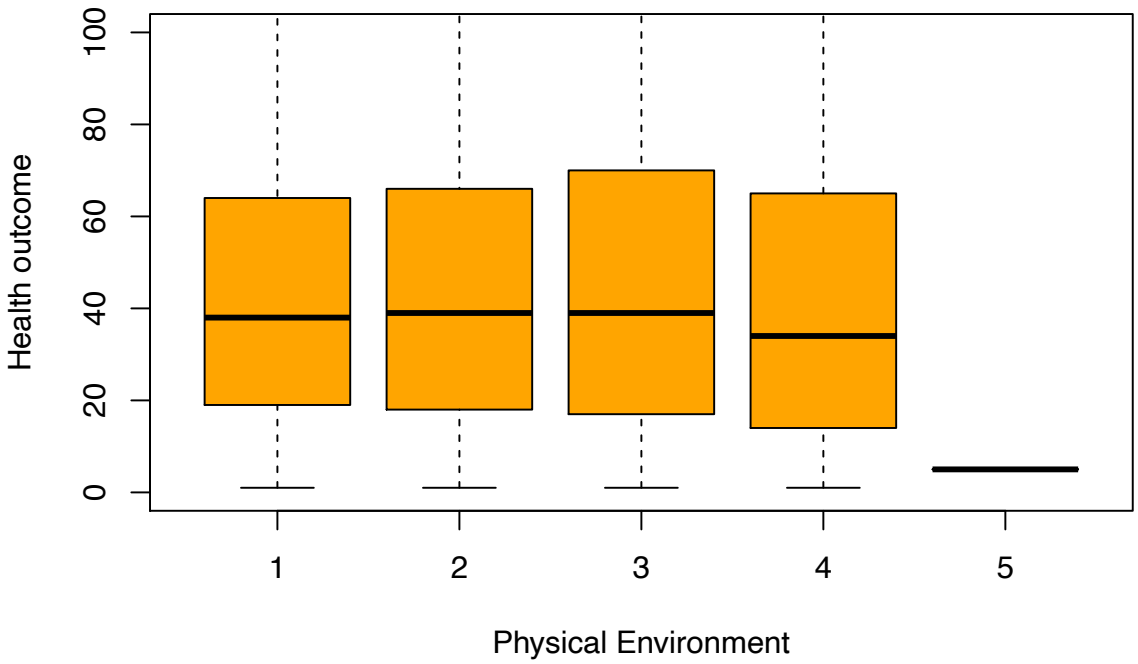
The method used in this study is mainly descriptive analysis and model building methods, the descriptive analysis is aimed to give an overview of relations among the predictors and the health outcome as there are too many predictors and the model building method used is a linear regression model. And in the model building part, firstly a full model with health outcome as response and all of the other factors used as predictors are included in the model at the same time, then a model selection method (Akaike’s information criterion [AIC]) backward selection method is applied to select a better subset of the model. And then model diagnostics would be performed on the selected model and data transformation is applied to obtain the final model, also,

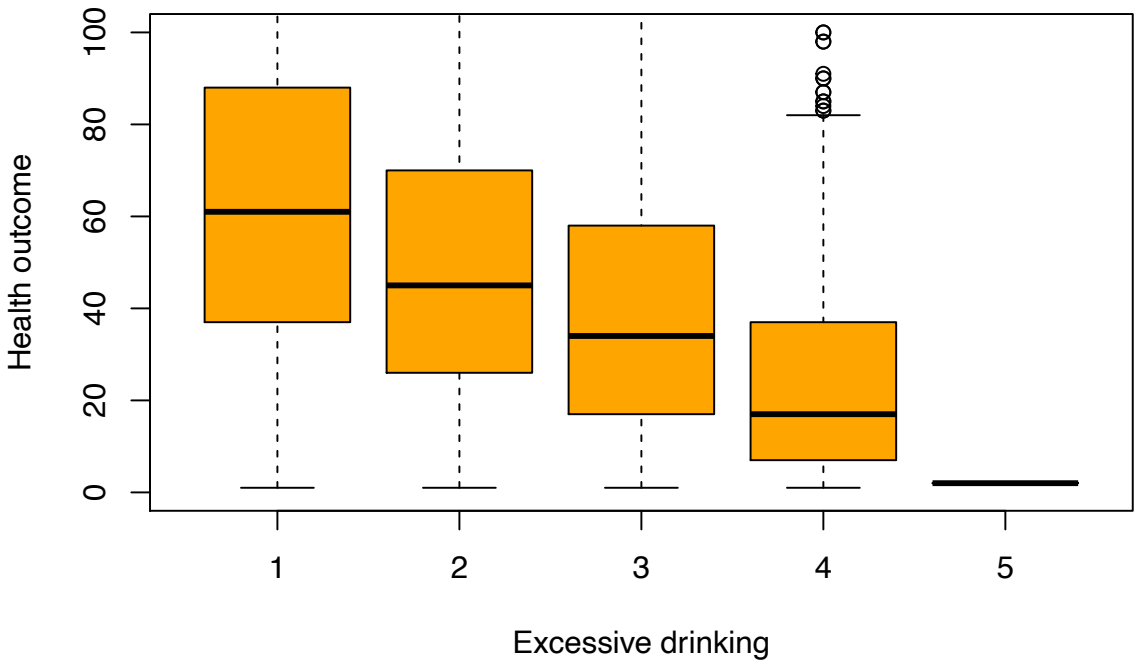
unusual points would be removed to improve the performances of the model, finally, with the results of the final model, inferences are made to explain which factors are most important ones for better health outcome.

Results

Descriptive analysis

After the data is cleaned, basic descriptive analysis mainly based on graphical summarises are firstly applied to give an overview of relations among the predictors and the health outcome, as there are too many predictors in the data, only two of them are selected as an example:





The above two boxplots show that there are different relations among different predictors and the health outcome, for factor Excessive drinking, it can be seen that the average levels of health outcomes are much different across the different levels of Excessive drinking but for factor Physical Environment, there are also differences but not so obviously. This indicates that a formal model building procedure is needed to study the relations between predictors and the health outcome better.

Models

As mentioned in previous parts, the response health outcome for this study is ranked data which is ordered numeric values, so the linear regression model is appropriate to modeling ranked data, especially, the rank range is from 1 to 242 which is a wide range. And the predictors are other measures data including Physical Environment, Adult obesity, and so on. As the predictors used are Quartile data which ranged from 1 to 5 and they are also ordered, so these predictors can also be treated as numeric. Firstly, the full linear model using health outcome rank as response and all of the other measures as predictors, the output is:

```
##
## Call:
## lm(formula = Y ~ ., data = cleandata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.990  -15.779   -3.929    7.872  169.670
##
## Coefficients: (1 not defined because of singularities)
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -17.17043    6.47635  -2.651 0.008061
## `Length of Life`                    13.26229    0.82965  15.985  < 2e-16
## `Quality of Life`                    5.27094    1.40840   3.743 0.000186
## `Health Behaviors`                   2.45614    1.34522   1.826 0.067975
## `Clinical Care`                     -0.33231    1.22000  -0.272 0.785343
## `Social & Economic Factors`          0.44102    1.38618   0.318 0.750387
## `Physical Environment`              -2.22648    1.04358  -2.134 0.032963
## `Premature death`                         NA         NA      NA       NA
```

```
## `Poor or fair health`                    -1.19248    1.19039  -1.002 0.316541
## `Poor physical health days`               1.84139    1.29357   1.423 0.154696
## `Poor mental health days`                 2.02744    1.00022   2.027 0.042750
## `Low birthweight`                         3.78083    0.88659   4.264 2.07e-05
## `Adult smoking`                           0.98545    1.14668   0.859 0.390192
## `Adult obesity`                          -1.73443    0.71729  -2.418 0.015663
## `Food environment index`                 -1.54866    0.69594  -2.225 0.026136
## `Physical inactivity`                    -1.23785    0.67834  -1.825 0.068126
## `Access to exercise opportunities`        0.12427    0.64786   0.192 0.847902
## `Excessive drinking`                     -1.90518    0.74337  -2.563 0.010428
## `Alcohol-impaired driving deaths`        -0.64497    0.55290  -1.167 0.243497
## `Sexually transmitted infections`        -0.62816    0.67878  -0.925 0.354822
## `Teen births`                             0.34037    0.78814   0.432 0.665872
## Uninsured                                -0.05321    0.81527  -0.065 0.947965
## `Primary care physicians`                 0.37623    0.70538   0.533 0.593818
## Dentists                                 -0.14057    0.68673  -0.205 0.837828
## `Mental health providers`                 0.81689    0.64489   1.267 0.205360
## `Preventable hospital stays`             -0.16774    0.75672  -0.222 0.824590
## `Diabetes monitoring`                    -0.22310    0.60704  -0.368 0.713248
## `Mammography screening`                   0.56184    0.64433   0.872 0.383287
## `High school graduation`                  0.65495    0.61785   1.060 0.289206
## `Some college`                           -0.24068    0.78787  -0.305 0.760015
## Unemployment                              0.27525    0.86377   0.319 0.750001
## `Children in poverty`                     0.98639    1.04545   0.944 0.345494
## `Income inequality`                       1.07559    0.64024   1.680 0.093063
## `Children in single-parent households`    0.27218    0.71428   0.381 0.703188
## `Social associations`                    -0.80500    0.59439  -1.354 0.175730
## `Violent crime`                           0.59002    0.63910   0.923 0.355980
## `Injury deaths`                          -0.76355    0.71261  -1.071 0.284041
## `Air pollution - particulate matter`      1.23468    0.75708   1.631 0.103029
## `Drinking water violations`               2.21124    0.85980   2.572 0.010164
## `Severe housing problems`                 0.15335    0.71512   0.214 0.830219
## `Driving alone to work`                   0.86423    0.66399   1.302 0.193158
## `Long commute - driving alone`            0.09364    0.66508   0.141 0.888039
##
## (Intercept)                             **
## `Length of Life`                        ***
## `Quality of Life`                       ***
## `Health Behaviors`                      .
## `Clinical Care`
## `Social & Economic Factors`
## `Physical Environment`                  *
## `Premature death`
## `Poor or fair health`
## `Poor physical health days`
## `Poor mental health days`               *
## `Low birthweight`                       ***
## `Adult smoking`
## `Adult obesity`                         *
## `Food environment index`                *
## `Physical inactivity`                   .
## `Access to exercise opportunities`
## `Excessive drinking`                    *
## `Alcohol-impaired driving deaths`
```

```
## `Sexually transmitted infections`
## `Teen births`
## Uninsured
## `Primary care physicians`
## Dentists
## `Mental health providers`
## `Preventable hospital stays`
## `Diabetes monitoring`
## `Mammography screening`
## `High school graduation`
## `Some college`
## Unemployment
## `Children in poverty`
## `Income inequality`                      .
## `Children in single-parent households`
## `Social associations`
## `Violent crime`
## `Injury deaths`
## `Air pollution - particulate matter`
## `Drinking water violations`              *
## `Severe housing problems`
## `Driving alone to work`
## `Long commute - driving alone`
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.29 on 3037 degrees of freedom
## Multiple R-squared:  0.4052, Adjusted R-squared:  0.3974
## F-statistic: 51.73 on 40 and 3037 DF,  p-value: < 2.2e-16
```

The model shows an R-squared value of 0.4052, and lots of predictors are insignificant, so the model is not appropriate, this might due to that there are lots of predictors that are highly correlated, include all of them in the model at the same time is not appropriate due to multi-collinearity. Thus, the model selection method is used to select a better subset of predictors, here, AIC backward selection method is used, the output is as below:

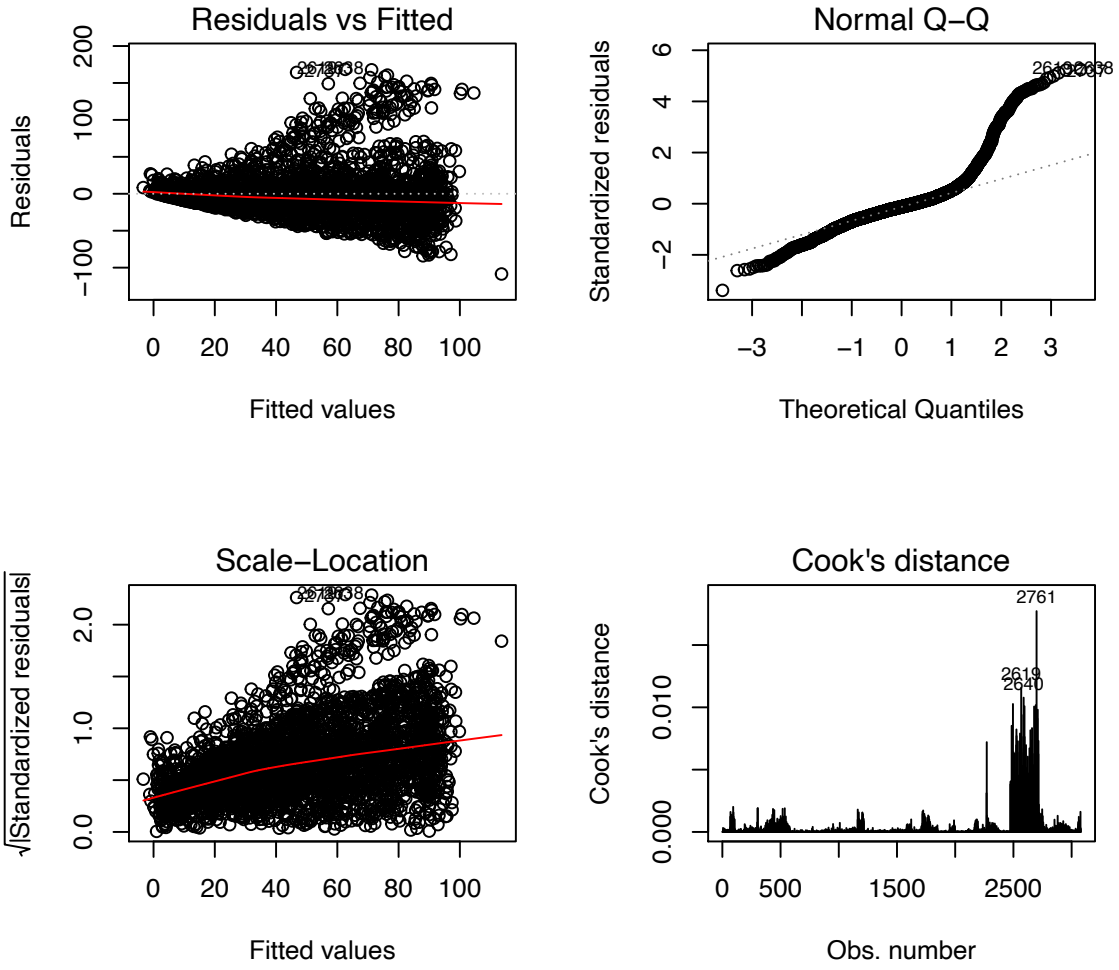
```
##
## Call:
## lm(formula = Y ~ `Length of Life` + `Quality of Life` + `Health Behaviors` +
##     `Physical Environment` + `Poor mental health days` + `Low birthweight` +
##     `Adult obesity` + `Food environment index` + `Physical inactivity` +
##     `Excessive drinking` + `Alcohol-impaired driving deaths` +
##     `Mental health providers` + `Children in poverty` + `Income inequality` +
##     `Air pollution - particulate matter` + `Drinking water violations` +
##     `Driving alone to work`, data = cleandata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.603  -15.768   -3.876    7.817  168.416
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -17.4766     4.7940  -3.646 0.000271 ***
## `Length of Life`                   13.2061     0.6930  19.056  < 2e-16 ***
## `Quality of Life`                   5.7834     1.2240   4.725  2.4e-06 ***
```

```
## `Health Behaviors`                     3.4007     0.9517   3.573 0.000358 ***
## `Physical Environment`                -2.1032     0.8651  -2.431 0.015108 *
## `Poor mental health days`              2.5950     0.9069   2.861 0.004246 **
## `Low birthweight`                      3.6126     0.8166   4.424  1.0e-05 ***
## `Adult obesity`                       -1.8468     0.6851  -2.696 0.007061 **
## `Food environment index`              -1.3630     0.6555  -2.079 0.037670 *
## `Physical inactivity`                 -1.3981     0.6495  -2.152 0.031437 *
## `Excessive drinking`                  -1.9661     0.7072  -2.780 0.005471 **
## `Alcohol-impaired driving deaths`     -0.7929     0.5378  -1.474 0.140525
## `Mental health providers`              0.8754     0.5572   1.571 0.116294
## `Children in poverty`                  1.4358     0.8896   1.614 0.106635
## `Income inequality`                    1.1632     0.6077   1.914 0.055687 .
## `Air pollution - particulate matter`   1.2175     0.7134   1.707 0.087985 .
## `Drinking water violations`            2.0920     0.7970   2.625 0.008716 **
## `Driving alone to work`                0.9881     0.5846   1.690 0.091087 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.23 on 3060 degrees of freedom
## Multiple R-squared:  0.4029, Adjusted R-squared:  0.3996
## F-statistic: 121.5 on 17 and 3060 DF,  p-value: < 2.2e-16
```

Now, it can be seen that the R-squared value is almost not changed, but lots of predictors are significant in the model compared with the full model, it means that this model is better. The VIF values below show that the model does not have any multicollinearity issue because all of VIF values are much less than 10.

```
##                                     vif.model2.
## `Length of Life`                       1.780805
## `Quality of Life`                      5.572803
## `Health Behaviors`                     3.369467
## `Physical Environment`                 2.784110
## `Poor mental health days`              3.059307
## `Low birthweight`                      2.480608
## `Adult obesity`                        1.746114
## `Food environment index`               1.631980
## `Physical inactivity`                  1.566686
## `Excessive drinking`                   1.860663
## `Alcohol-impaired driving deaths`      1.073469
## `Mental health providers`              1.165608
## `Children in poverty`                  2.949703
## `Income inequality`                    1.373676
## `Air pollution - particulate matter`   1.899691
## `Drinking water violations`            1.580471
## `Driving alone to work`                1.271248
```

The model diagnostics plots are as below:



Check model assumptions:

1) Independent assumption: the points are randomly distributed around the zero-mean line, the assumption is true.

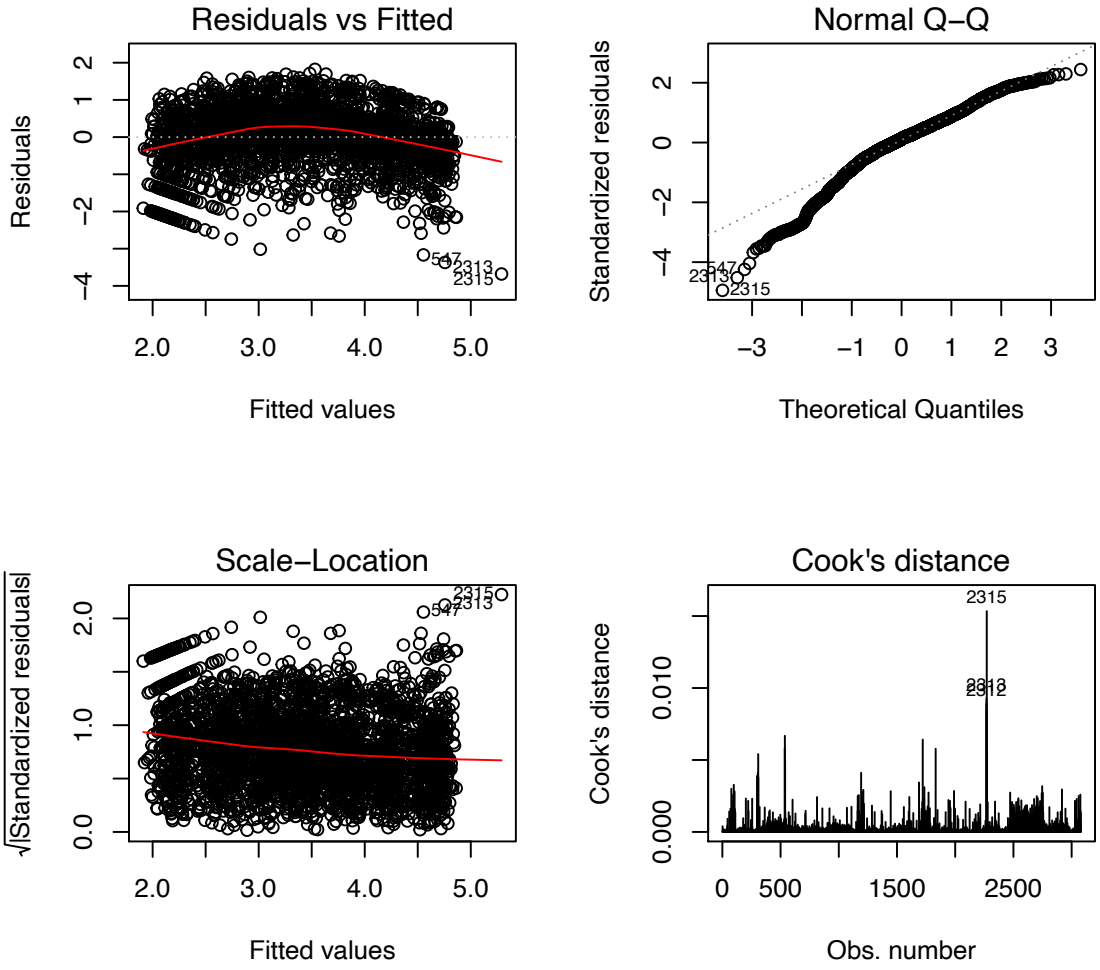
2) Linearity assumption: the residuals plot shows there is no special curve, the assumption is true.

3) Constant variance assumption: the residuals plot shows the spread of residuals change clearly across the x-axis that it becomes larger and larger; the assumption is not true.

4) Normality assumption: the normal Q-Q plot shows that the residuals do not fit the straight line well, the normality assumption is not true.

5) Unusual data points: The residuals plot and normal Q-Q plot show there are some points with relatively large absolute values of residuals, these points are supposed to be outliers. The Cook's distance plot shows there are observations with large Cook's distance, these points are unusual data points.

Thus, to obtain a better model, model transformation is applied and a log transformation of response is used, the result is:



Check model assumptions again, clearly, independent assumption and linearity assumption are true. And for constant variance assumption, the residuals plot shows the spread of residuals does not change across the x-axis, the assumption is true. And for normality assumption the normal Q-Q plot shows that the residuals fit the straight line well overall, the normality assumption is true. Thus, the model assumptions are true. However, there are still some unusual data, after deleting these points, the final model is as below:

```
##
## Call:
## lm(formula = log(Y) ~ `Length of Life` + `Quality of Life` +
##     `Health Behaviors` + `Physical Environment` + `Poor mental health days` +
##     `Low birthweight` + `Adult obesity` + `Food environment index` +
##     `Physical inactivity` + `Excessive drinking` + `Alcohol-impaired driving deaths` +
##     `Mental health providers` + `Children in poverty` + `Income inequality` +
##     `Air pollution - particulate matter` + `Drinking water violations` +
##     `Driving alone to work`, data = cleandata[-pos, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81165 -0.33885  0.03923  0.38654  1.44851
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.6574706  0.0934800  17.731  < 2e-16 ***
```

```
## `Length of Life`                      0.3543878  0.0133022  26.641  < 2e-16 ***
## `Quality of Life`                     0.2176367  0.0237312   9.171  < 2e-16 ***
## `Health Behaviors`                    0.0673026  0.0181556   3.707 0.000214 ***
## `Physical Environment`               -0.0509728  0.0167185  -3.049 0.002318 **
## `Poor mental health days`             0.0596739  0.0175109   3.408 0.000664 ***
## `Low birthweight`                     0.0693499  0.0157884   4.392 1.16e-05 ***
## `Adult obesity`                      -0.0008846  0.0129731  -0.068 0.945639
## `Food environment index`             -0.0167493  0.0126464  -1.324 0.185464
## `Physical inactivity`                 0.0248944  0.0123562   2.015 0.044027 *
## `Excessive drinking`                 -0.0346815  0.0137035  -2.531 0.011432 *
## `Alcohol-impaired driving deaths`    -0.0140990  0.0102920  -1.370 0.170825
## `Mental health providers`             0.0140610  0.0106576   1.319 0.187160
## `Children in poverty`                 0.0158960  0.0172245   0.923 0.356150
## `Income inequality`                  -0.0077552  0.0117174  -0.662 0.508118
## `Air pollution - particulate matter`  0.0083750  0.0136774   0.612 0.540374
## `Drinking water violations`           0.0395228  0.0152053   2.599 0.009390 **
## `Driving alone to work`               0.0085655  0.0112433   0.762 0.446225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5949 on 2864 degrees of freedom
## Multiple R-squared:  0.6046, Adjusted R-squared:  0.6023
## F-statistic: 257.6 on 17 and 2864 DF,  p-value: < 2.2e-16
```

Thus, finally, it can be seen that the adjusted R-squared improved a lot which is now over 0.60 which means there is over 60% variation in the health outcome that could be explained by the model now. The goodness of fit is not bad. And from the output of the model, it can be seen that the factors Length of Life, Quality of Life, Health Behaviors, Physical Environment, Poor mental health days, Low birthweight, Physical inactivity, Excessive drinking, and Drinking water violations are the key factors which have significant effects on the health outcomes at 5% significance level.

Conclusion

After all of the above work, with the best optimal model selected by AIC backward method, the model diagnostics show transformation is needed and after using log transformation, the model diagnostics show all of the linear regression assumptions are true that the model is valid and the inferences based on the data is reliable. And to improve the final model, the unusual points are removed too, the final model shows not bad goodness of fit and the final model factors Length of Life, Quality of Life, Health Behaviors, Physical Environment, Poor mental health days, Low birthweight, Physical inactivity, Excessive drinking, and Drinking water violations are the key factors for better health outcome. For example, each additional level increase in the Physical Environment is associated with 0.0509728 units decrease in the rank of the health outcome indicating better health outcomes (lower rank, better health). At last, besides all of the above findings, there are also some limitations, for examples, only linear regression model is used in the study, other models such as logistic model, poisson model might be more appropriate, also, the predictors used are in quartile scale, other scales might give better results, all of these issues could be investigated in future study to obtain better results.

Bibliography

Kevin, K. (2018). Report Shows Health Disparities by County, Demographics. https://www.aafp.org/news/health-of-the-public/20180327disparitiesrpt.html

RWJF. (2018). 2018 County Health Rankings Key Findings Report. https://www.countyhealthrankings.org/

Hamad, R., Brown, D.M. & Basu, S. (2019) The association of county-level socioeconomic factors with individual tobacco and alcohol use: a longitudinal study of U.S. adults. BMC Public Health 19, 390. https://doi.org/10.1186/s12889-019-6700-x

Wallace, M., Sharfstein, J. M., Kaminsky, J., & Lessler, J. (2019). Comparison of US County-Level Public Health Performance Rankings With County Cluster and National Rankings: Assessment Based on Prevalence Rates of Smoking and Obesity and Motor Vehicle Crash Death Rates. JAMA network open, 2(1), e186816. https://doi.org/10.1001/jamanetworkopen.2018.6816

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Kutz, M. J., Huynh, C., Barber, R. M., Shackelford, K. A., Mackenbach, J. P., van Lenthe, F. J., Flaxman, A. D., Naghavi, M., Mokdad, A. H., & Murray, C. J. (2016). US County-Level Trends in Mortality Rates for Major Causes of Death, 1980-2014. JAMA, 316(22), 2385–2401. https://doi.org/10.1001/jama.2016.13645

Stone, L. C., Boursaw, B., Bettez, S. P., Larzelere Marley, T., & Waitzkin, H. (2015). Place as a predictor of health insurance coverage: A multivariate analysis of counties in the United States. Health & place, 34, 207–214. https://doi.org/10.1016/j.healthplace.2015.03.015

Roth, G. A., Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Naghavi, M., Mokdad, A. H., & Murray, C. (2017). Trends and Patterns of Geographic Variation in Cardiovascular Mortality Among US Counties, 1980-2014. JAMA, 317(19), 1976–1992. https://doi.org/10.1001/jama.2017.4150

Garcia, M. C., Rossen, L. M., Bastian, B., Faul, M., Dowling, N. F., Thomas, C. C., Schieb, L., Hong, Y., Yoon, P. W., & Iademarco, M. F. (2019). Potentially Excess Deaths from the Five Leading Causes of Death in Metropolitan and Nonmetropolitan Counties - United States, 2010-2017. Morbidity and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002), 68(10), 1–11. https://doi.org/10.15585/mmwr.ss6810a1

MACPAC (Medicaid and CHIP Payment and Access Commission). 2016. Medicaid access in brief: Adults' experiences in obtaining medical care. Washington, DC: MACPAC. https://www.macpac.gov/wp-content/uploads/2016/11/Adults-Experiences-in-Obtaining-Medical-Care.pdf

Singh, G. K., Daus, G. P., Allender, M., Ramey, C. T., Martin, E. K., Perry, C., Reyes, A., & Vedamuthu, I. P. (2017). Social Determinants of Health in the United States: Addressing Major Health Inequality Trends for the Nation, 1935-2016. International journal of MCH and AIDS, 6(2), 139–164. https://doi.org/10.21106/ijma.236

Appendix

Linear model

The linear model used in the study has the following form:

$$Y = \beta_0 + \beta_1 x_1 + ...\beta_p x_p + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$, i.i.d.

AIC model selection

The model selection method used is AIC backward model selection which uses the formula:

$$AIC = -2(log - likelihood) + 2K$$

where K is the number of model parameters including the intercept and the lower AIC the better the linear model. The AIC backward model selection starts from the full model and delete predictors until the lowest AIC is achieved that no predictors could be removed.

R codes

```
`
library(readxl)
data1 <- read_excel("2018 County Health Rankings Data - v2.xls", sheet = 2)
data2 <- read_excel("2018 County Health Rankings Data - v2.xls", sheet = 3)
data3 <- read_excel("2018 County Health Rankings Data - v2.xls", sheet = 4)

nms1 <- colnames(data1)
nms2 <- colnames(data2)
nms3 <- colnames(data3)
data1 <- read_excel("2018 County Health Rankings Data - v2.xls", sheet = 2, skip = 1, na = "NR")
data2 <- read_excel("2018 County Health Rankings Data - v2.xls", sheet = 3, skip = 1)
data3 <- read_excel("2018 County Health Rankings Data - v2.xls", sheet = 4, skip = 1)

data2 <- data2[,c(1:3,which(grepl("Quartile",colnames(data2))))]
data3 <- data3[,c(1:3,which(grepl("Quartile",colnames(data3))))]


data1 <-  data1[,c(1,2,3,5)]
colnames(data1)[4] <- "Y"


colnames(data2)[-c(1:3)] <-  nms2[!grepl("[1-9]",nms2)]
colnames(data3)[-c(1:3)] <-  nms3[!grepl("[1-9]",nms3)]

m1 <- merge(data1, data2, by = c("FIPS","State","County"))
m2 <- merge(m1, data3, by = c("FIPS","State","County"))

cleandata <- na.omit(m2)[,-c(1:3)]

for(i in 1:ncol(cleandata)) {
  cleandata[,i] <- as.numeric(as.character(cleandata[,i] ))
}
```

```r
par(mfrow = c(1,1))
boxplot(Y ~ `Physical Environment`, data = cleandata, col="orange", ylim = c(0,100),
        ylab = "Health outcome")

par(mfrow = c(1,1))
boxplot(Y ~ `Excessive drinking`, data = cleandata, col="orange", ylim = c(0,100),
        ylab = "Health outcome")

model1 <- lm(Y ~ ., data = cleandata)
summary(model1)

model2 <- step(model1, direction = "backward", trace = 0)
summary(model2)

library(car)
data.frame(vif(model2))

par(mfrow = c(2,2))
plot(model2, 1:4)

model3 <- lm(log(Y) ~ `Length of Life` + `Quality of Life` + `Health Behaviors` +
    `Physical Environment` + `Poor mental health days` + `Low birthweight` +
    `Adult obesity` + `Food environment index` + `Physical inactivity` +
    `Excessive drinking` + `Alcohol-impaired driving deaths` +
    `Mental health providers` + `Children in poverty` + `Income inequality` +
    `Air pollution - particulate matter` + `Drinking water violations` +
    `Driving alone to work`, data = cleandata)

par(mfrow = c(2,2))
plot(model3, 1:4)

pos <- which(cooks.distance(model3) > 4/(nrow(cleandata) - length(coef(model3))) | abs(rstudent(model3))
model4 <- lm(log(Y) ~ `Length of Life` + `Quality of Life` + `Health Behaviors` +
    `Physical Environment` + `Poor mental health days` + `Low birthweight` +
    `Adult obesity` + `Food environment index` + `Physical inactivity` +
    `Excessive drinking` + `Alcohol-impaired driving deaths` +
    `Mental health providers` + `Children in poverty` + `Income inequality` +
    `Air pollution - particulate matter` + `Drinking water violations` +
    `Driving alone to work`, data = cleandata[-pos,])

summary(model4)
```